Dual of Logistic regression.

Recall logistic regression:

Data points \( (\mathbf{x}_i) \), labels \( y_i = +1 \) or \(-1\).

Want: \( \mathbf{w}^T \mathbf{x} + \beta = \log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} \), where \( p(\mathbf{x}) \) is the probability that the data point belongs to class 1.

We showed that finding the best \( \mathbf{w}, \beta \) is the same as:

Maximize \( \prod_{i=1}^{n} \left( \frac{\exp (y_i; (\mathbf{w}^T \mathbf{x}_i + \beta))}{1 + \exp (y_i; (\mathbf{w}^T \mathbf{x}_i + \beta))} \right) \) (per the maximum-likelihood estimator)

Non-convex problem.

We can take logs to turn this into a convex problem.

Maximize \( \log \left( \prod_{i=1}^{n} \left( \frac{\exp (y_i; (\mathbf{w}^T \mathbf{x}_i + \beta))}{1 + \exp (y_i; (\mathbf{w}^T \mathbf{x}_i + \beta))} \right) \)\)

Maximize \( \sum_{i=1}^{n} \log \left( \frac{1}{1 + \exp (-y_i; (\mathbf{w}^T \mathbf{x}_i + \beta))} \right) \)

Maximize \( \sum_{i=1}^{n} -\log \left( 1 + \exp (y_i; (\mathbf{w}^T \mathbf{x}_i + \beta)) \right) \)
Instead of \( \max_x f(x) \), we consider \( -\min_x f(x) \).

The solution to the latter also solves the former.

So consider:

\[
p^* = \min_{w, \beta} \sum_{i=1}^{n} \log \left( 1 + \exp \left( y_i (\mathbf{w}^T \mathbf{x}_i + \beta) \right) \right)
\]

This is a convex problem.

It is unconstrained? What is its dual?

Let \( f(t) = \log(1 + e^{-t}) \).

Consider \( \beta = 0 \) for simplicity.

\[
p^* = \min_{\mathbf{w}} \sum_{i=1}^{n} \log \left( 1 + \exp \left( -y_i (\mathbf{w}^T \mathbf{x}_i) \right) \right)
\]

\[
= \min_{\mathbf{w}} \sum_{i=1}^{n} \log \left( 1 + \exp \left( -y_i \mathbf{x}_i^T \mathbf{w} \right) \right)
\]

Define \( \mathbf{A} = [y_1 \mathbf{x}_1, y_2 \mathbf{x}_2, \ldots, y_n \mathbf{x}_n] \) and \( \mathbf{v} = \mathbf{A}^T \mathbf{w} \).

\[
= \min_{\mathbf{v}, \mathbf{w}} \sum_{i=1}^{n} \log \left( 1 + \exp \left( -v_i \right) \right)
\]

\[
s.t. \mathbf{v} = \mathbf{A} \mathbf{w}
\]

Now we have a "constraint".
\[ p^* = \min_{\vec{v}, \vec{w}} \sum_{i=1}^{n} f(v_i). \quad \text{Convex.} \]

- Only constraints are linear equality + convex problem.
- If there is a feasible point, i.e. some point \( \vec{v} = A\vec{w} \), then Slater's condition holds, i.e. \( p^* = d^* \).

What is \( d^* \)?

\[ L(\vec{v}, \vec{w}, \vec{z}) = \sum_{i=1}^{n} f(v_i) + \vec{z}^T (\vec{v} - A\vec{w}) \]

\[ g(\vec{z}) = \min_{\vec{v}, \vec{w}} L(\vec{v}, \vec{w}, \vec{z}) = \min_{\vec{v}} \min_{\vec{w}} L(\vec{v}, \vec{w}, \vec{z}) \]

\[ = \min_{\vec{v}} \left\{ \begin{array}{l} -\infty \quad \text{if } A\vec{z} \neq 0. \\ \sum_{i=1}^{n} f(v_i) + \vec{z}^T \vec{v} \quad \text{if } A\vec{z} = 0. \end{array} \right. \]

Consider \( h(t) = f(t) + 2v_i t \)

\[ \frac{dh}{dt} = \frac{1}{1+e^{-t}} (e^{-t})(-1) + 2v_i = \frac{-1}{1+e^{t}} + 2v_i. \]

Setting \( = 0 \)

\[ \frac{1}{1+e^{t}} = 2v_i \quad \Rightarrow \quad 1+e^{t} = \frac{1}{2v_i} \]

\[ \Rightarrow e^{t} = \frac{1-2v_i}{2v_i} \quad \Rightarrow \quad t = \log \left( \frac{1-2v_i}{2v_i} \right). \]

Note \( \frac{1-2v_i}{2v_i} \geq 0 \) only when \( 2v_i \in [0, 1] \).

If \( 2v_i < 0 \), \( \lim_{t \to \infty} h(t) = -\infty \).

If \( 2v_i > 1 \), since \( \frac{1}{dt} \log(1+e^{-t}) = \frac{-1}{1+e^{t}} \) has slope \( < -1 \),

\[ 2v_i t \text{ dominates as } t \to -\infty \Rightarrow \lim_{t \to -\infty} h(t) = -\infty. \]
So for $y_i \in \{0,1\}$,

$$e^t = \frac{1-2y_i}{2}, \quad t = \log \left( \frac{1-2y_i}{2} \right).$$

$$h(y_i) = \log \left( 1 + e^{-\log \left( \frac{1-2y_i}{2} \right)} \right) + 2y_i \left( \log \left( \frac{1-2y_i}{2} \right) \right).$$

$$= \log \left( 1 + e^{\log \left( \frac{1-2y_i}{2} \right)} \right) + 2y_i \left( \log \left( \frac{1-2y_i}{2} \right) \right).$$

$$= \log \left( \frac{1}{1-2y_i} \right) + 2y_i \log \left( \frac{1}{2} \right) + 2y_i \cdot \log(1-2y_i).$$

$$= -2y_i \log 2y_i = (1-2y_i) \log (1-2y_i)$$

$$= \text{entropy of } 2y_i$$

$$g(\mathbb{z}) = \sum_{i=1}^{n} -2y_i \log 2y_i - (1-2y_i) \log (1-2y_i)$$

if $\mathbb{z}_i = 0$ and $y_i \in \{0,1\}$, $\forall i$

otherwise.

**Dual:**

$$\max_{\mathbb{z}} \sum_{i=1}^{n} \frac{1}{2y_i - (1-2y_i) \log (1-2y_i)}$$

subject to $y_i \in \{0,1\}, \forall i$ and $\mathbb{z}_i = 0$.

What is $2y_i$?

$$\frac{1}{p(y_i = y_i)} = \frac{1}{1-2y_i}$$

So $2y_i = P(Y_i \neq y_i)$

$A\mathbb{z} = 0 \Rightarrow \{x_1 \cdots x_n\} \begin{bmatrix} y_1 \cdot \left( 1 - p(y_i = y_i) \right) \\ y_n \cdot \left( 1 - p(y_n = y_n) \right) \end{bmatrix}$

Note $y_i \cdot (1 - p(y_i = y_i)) = 1 - p(\mathbb{z})$ if $y_i = 1$

$= p(\mathbb{z})$ if $y_i = -1$. 
Total Least Squares

In a normal least squares problem (sometimes called OLS for ordinary least squares) we try to find \( \hat{x} \) such that

\[
A \hat{x} \approx \hat{b}
\]

and we minimize

\[
||\hat{e}||_2^2 = ||A \hat{x} - \hat{b}||_2^2
\]

In this formulation, we are assuming that the errors in our data are only in \( \hat{b} \). But what if there are also errors in \( A \)?

We have

\[
(A + \tilde{A}) \hat{x} = \tilde{b} + \tilde{b}
\]

\[
\text{error in } A
\]

\[
\text{error in } \hat{b}
\]

\( \tilde{A} \) and \( \tilde{b} \) are perturbations we do not know.

We wish to find \( \hat{x} \) s.t. \( (A + \tilde{A}) \hat{x} = \tilde{b} + \tilde{b} \) but all we have is \( A \) and \( \hat{b} \).

How to measure the perturbations? And minimize them? i.e. find \( A, \hat{b} \) closest to \( A + \tilde{A}, \tilde{b} + \tilde{b} \).

Minimize

\[
||\tilde{A} \tilde{b}||_F
\]

subject to

\[
(A + \tilde{A}) \hat{x} = \tilde{b} + \tilde{b}
\]

\[
[A + \tilde{A}] \hat{x} = \tilde{b} + \tilde{b}
\]

\[
[A + \tilde{A}]\hat{x} = \tilde{b} + \tilde{b}
\]

\[
[A + \tilde{A}]\hat{x} = \tilde{b} + \tilde{b}
\]

\[
[A + \tilde{A}]\hat{x} = \tilde{b} + \tilde{b}
\]

So we want

\[
\begin{bmatrix} \hat{x} \\ -1 \end{bmatrix} \in N\left( \begin{bmatrix} A + \tilde{A} \\ \tilde{b} + \tilde{b} \end{bmatrix} \right)
\]

So this matrix must be rank deficient.
Define: $[A+\tilde{A} | \tilde{B} + \tilde{B}] = \tilde{Z} \in \mathbb{R}^{m \times (n+1)}$

$\tilde{A} \begin{bmatrix} A \end{bmatrix} = Z \in \mathbb{R}^{m \times (n+1)}$

minimize $\|Z - \tilde{Z}\|_F^2$

s.t. $\text{Rank}(\tilde{Z}) = n$

$\rightarrow$ solution by Eckart-Young.